УДК 655. 244. 07

М.С.Петрик

ЧАСТОТА ЗУСТРІЧІ ЗНАКІВ В УКРАЇНОМОВНИХ ТЕКСТАХ

Ймовірність зустрічі окремих знаків визначають статистичними методами аналізу текстових масивів. Літературні джерела [1, 2, 3] подають значення частоти зустрічі знаків (ЧЗЗ), проте вони стосуються переважно російськомовних текстів, не враховують стилю тексту і типу видання. Окрім того, у них ігнорується частина знакового асортименту шрифтів — великі літери, спеціальні символи, цифри.

Наведені недоліки, а також деякі зміни, що відбулися в граматиці, лексиці і стилістиці сучасної української мови, спонукають до проведення більш ретельного аналізу ЧЗЗ, визначають актуальність і необхідність досліджень.

ЧЗЗ у виданні визначали статистичним аналізом текстового масиву за допомогою програми, написаної на мові QuickBasic (Miccrosoft Copr.). Існуючі програми підрахунку появи окремих знаків [2] виконують такі кроки, як: перебір кожного знака текстового масиву (крок 1); заявлення масиву знаків, які підлягають аналізу (крок 2); збільшення значення лічильника на одиницю при збіжності знака з шуканим (крок 3).

```
L = LEN(S$)                                                    , крок 1
IF L > 0 THEN
FOR I = 1 TO L
  SYM$ = MID$(S$, I, 1)
  FOR J = 32 TO 249                                            , крок 2
IF (J>63 AND J<128) OR (J>175 AND J<224) GOTO MITKA
IF SYM$ = CIIR$(J) THEN QuanSym(J) = QuanSym(J) + 1  , крок 3
MITKA:
  NEXT J
NEXT I
END IF
```

Такий аналіз текстового масиву є трудомістким, оскільки вимагає кількості перевірок кожного знака, рівної величині заявленого масиву. Пропонується інший алгоритм, який полягає в рекурсивному виклику масиву знаків: перебір кожного знака текстового масиву (крок 4); виклик з масиву знаків лічильника знайденого знака, який збільшує своє значення на одиницю (крок 5).

```
L = LEN(S$)                                          , крок 4
IF L > 0 THEN
  FOR I = 1 TO L
      SYM$ = MID$(S$, I, 1) . J = ASC(SYM$)
      QuanSym(J) = QuanSym(J) + 1                    , крок 5
  NEXT I
END IF
```

Робота програми згідно з цим алгоритмом значно прискорюється, тому що відсутня операція перевірки знака на відповідність заявленому масиву.

Аналізу підлягали п'ять типів поліграфічних видань, записаних на магнітних носіях інформації, які характеризувались різним стилем текстів і розмір яких перевищував 500 кілобайт. Результати досліджень подані в таблиці.

Код	Тип видання					
	наук. пуб. ж.	наук. ж.	худ. кн.	наук. т. кн.	публ. кн.	літер.
Частота						
< >	0.13020	0.12875	0.14745	0.11727	0.12391	.
<!>	0.00004	0.00021	0.00235	0.00000	0.00006	.
<">	0.00198	0.00752	0.00096	0.00247	0.00131	.
<-->	0.00324	0.00347	0.00353	0.00332	0.00015	.
<%>	0.00016	0.00000	0.00000	0.00014	0.00000	.
<..	0.00006	0.00038	0.00004	0.00003	0.00014	.
<'>	0.00185	0.00108	0.00002	0.00111	0.00151	.
<(>	0.00131	0.00177	0.00005	0.00151	0.00189	.
<)>	0.00136	0.00181	0.00006	0.00158	0.00191	.

Код	Тип видання					
	наук. пуб. ж.	наук. ж.	худ. кн.	наук. т. кн.	публ. кн.	літер.
<*>	0.00006	0.00016	0.00000	0.00000	0.00001	.
<+>	0.00018	0.00001	0.00000	0.00001	0.00000	.
<,>	0.01322	0.01556	0.02030	0.01341	0.01508	.
<->	0.00111	0.00140	0.00726	0.00189	0.00537	.
<.>	0.01252	0.01457	0.01371	0.01418	0.01292	.
</>	0.00007	0.00066	0.00001	0.00007	0.00009	.
<0>	0.00203	0.00101	0.00003	0.00147	0.00031	.
<1>	0.00205	0.00377	0.00006	0.00219	0.00088	.
<2>	0.00159	0.00171	0.00005	0.00122	0.00040	.
<3>	0.00129	0.00124	0.00002	0.00080	0.00020	.
<4>	0.00089	0.00103	0.00002	0.00075	0.00020	.
<5>	0.00101	0.00093	0.00002	0.00104	0.00021	.
<6>	0.00043	0.00084	0.00001	0.00072	0.00016	.
<7>	0.00056	0.00087	0.00003	0.00051	0.00023	.
<8>	0.00049	0.00150	0.00002	0.00068	0.00026	.
<9>	0.00071	0.00204	0.00005	0.00075	0.00041	.
<:>	0.00053	0.00110	0.00086	0.00057	0.00049	.
<;>	0.00035	0.00058	0.00006	0.00048	0.00014	.
<< >	0.00000	0.00011	0.00001	0.00020	0.00000	.
<=>	0.00002	0.00002	0.00000	0.00001	0.00000	.
<>>	0.00000	0.00011	0.00001	0.00020	0.00000	.
<?>	0.00019	0.00013	0.00091	0.00078	0.00005	.
<А>	0.00123	0.00096	0.00052	0.00056	0.00185	.
<Б>	0.00058	0.00087	0.00085	0.00053	0.00084	.
<В>	0.00173	0.00216	0.00213	0.00206	0.00229	.
<Г>	0.00050	0.00113	0.00038	0.00054	0.00078	

Код	Тип видання					
	наук. пуб. ж.	наук. ж.	худ. кн.	наук. т. кн.	публ. кн.	літер.
<Д>	0.00101	0.00108	0.00232	0.00114	0.00090	
<Е>	0.00051	0.00022	0.00004	0.00039	0.00042	
<Ж>	0.00013	0.00017	0.00027	0.00011	0.00035	
<З>	0.00065	0.00080	0.00111	0.00082	0.00112	
<И>	0.00055	0.00018	0.00003	0.00029	0.00061	
<Й>	0.00024	0.00021	0.00009	0.00017	0.00013	
<К>	0.00116	0.00142	0.00084	0.00135	0.00247	
<Л>	0.00085	0.00129	0.00036	0.00065	0.00080	
<М>	0.00085	0.00171	0.00049	0.00096	0.00071	
<Н>	0.00168	0.00112	0.00144	0.00128	0.00197	
<О>	0.00131	0.00101	0.00116	0.00093	0.00140	
<П>	0.00206	0.00188	0.00233	0.00188	0.00235	
<Р>	0.00093	0.00099	0.00029	0.00098	0.00120	
<С>	0.00127	0.00212	0.00126	0.00137	0.00167	
<Т>	0.00128	0.00127	0.00124	0.00089	0.00125	
<У>	0.00122	0.00126	0.00044	0.00115	0.00151	
<Ф>	0.00019	0.00075	0.00005	0.00021	0.00013	
<Х>	0.00025	0.00088	0.00034	0.00088	0.00042	
<Ц>	0.00048	0.00036	0.00026	0.00034	0.00053	
<Ч>	0.00047	0.00027	0.00047	0.00031	0.00072	
<Ш>	0.00014	0.00090	0.00005	0.00024	0.00031	
<Щ>	0.00017	0.00032	0.00035	0.00013	0.00007	
<Ь>	0.00013	0.00007	0.00002	0.00007	0.00015	
<Ю>	0.00006	0.00018	0.00011	0.00015	0.00004	
<Я>	0.00083	0.00054	0.00075	0.00108	0.00043	
<а>	0.05881	0.06049	0.07277	0.05796	0.06256	0.0736

Код	Тип видання					
	наук. пуб. ж.	наук. ж.	худ. кн.	наук. т. кн.	публ. кн.	літер.
<б>	0.01440	0.00918	0.01641	0.01535	0.01389	0.0171
<в>	0.04130	0.04004	0.04610	0.04225	0.04149	0.0523
<г>	0.01051	0.01269	0.01393	0.01101	0.01292	0.0147
<д>	0.02835	0.02261	0.02895	0.02266	0.02259	0.0293
<е>	0.03578	0.03726	0.03778	0.03712	0.02980	0.0423
<ж>	0.00886	0.00652	0.00743	0.00600	0.00619	0.0078
<з>	0.01726	0.01410	0.01742	0.01724	0.01785	0.0209
<и>	0.05436	0.04690	0.05403	0.05399	0.05954	0.0607
<й>	0.00815	0.00950	0.01294	0.00813	0.01020	0.0112
<к>	0.03297	0.02964	0.02850	0.03455	0.03827	0.0346
<л>	0.02923	0.02651	0.03229	0.02970	0.03352	0.0356
<м>	0.02455	0.02112	0.01947	0.02532	0.02328	0.0268
<н>	0.05502	0.05227	0.04084	0.06104	0.05931	0.0617
<о>	0.07578	0.07323	0.07142	0.07964	0.08446	0.0908
<п>	0.02141	0.01929	0.02129	0.01900	0.01954	0.0272
<р>	0.03912	0.03774	0.03261	0.04813	0.03891	0.0456
<с>	0.03136	0.03321	0.03279	0.03002	0 03218	0.0387
<т>	0.04011	0.03879	0.03767	0.04455	0.03391	0.0448
<у>	0.02549	0.02477	0.02819	0.02583	0.02665	0.0302
<ф>	0.00150	0.00239	0.00046	0.00310	0.00144	0.0012
<х>	0.01087	0.00805	0.00913	0.01016	0.01046	0.0106
<ц>	0.00559	0.00689	0.00428	0.00606	0.00564	0.0086
<ч>	0.00978	0.01125	0.00875	0.00936	0.01260	0.0118
<ш>	0.00554	0.00502	0.00839	0.00568	0.00840	0.0087
<щ>	0.00529	0.00295	0.00551	0.00309	0.00284	0.0043
<ь>	0.01413	0.01510	0.01073	0.01759	0.01222	0.0177

Код	Тип видання					
	наук. пуб. ж.	наук. ж.	худ. кн.	наук. т. кн.	публ. кн.	літер.
<ю>	0.00758	0.00592	0.00587	0.01118	0.00668	0.0080
<я>	0.01684	0.01503	0.01732	0.01836	0.01864	0.0200
<Г>	0.00000	0.00000	0.00008	0.00000	0.00000	.
<г>	0.00000	0.00000	0.00069	0.00000	0.00000	.
<Є>	0.00013	0.00021	0.00003	0.00007	0.00008	.
<є>	0.00437	0.00471	0.00368	0.00390	0.00257	0.0001
<І>	0.00087	0.00143	0.00083	0.00077	0.00079	.
<і>	0.05053	0.04418	0.04416	0.04481	0.04692	0.0570
<І>	0.00010	0.00005	0.00011	0.00000	0.00017	.
<ї>	0.00440	0.00702	0.00349	0.00469	0.00567	0.0074
Сума частот						
	0.992269	0.966492	0.994144	0.999104	0.997981	

Результати розрахунків (див. таблицю) засвідчують певні відхилення в значеннях ЧЗЗ порівняно з літературними даними (остання колонка таблиці). Це пояснюється тим, що запропонований масив знаків для підрахунку є більш коректним, оскільки текстові файли, які підлягають комп'ютерній обробці, містять розширений асортимент знаків.

Як і очікувалось, стиль текстового масиву значною мірою впливає на ЧЗЗ. Так, розділові знаки, що більш характерні для художніх видань, мають вище значення ЧЗЗ. Цифри і математичні символи частіше зустрічаються в науково-технічних виданнях. Цікавим є той факт, що серед цифр найбільше значення ЧЗЗ в "1", найменше — у "6" і "7".

Отримані результати дозволяють розрахувати математично-очікувану ширину знаків комп'ютерних шрифтів і побудувати таблиці місткості шрифтів, а це, безперечно,

важливо для сучасної технології поліграфічної обробки текстів. Отримані дані дозволяють розв'язати обернену задачу, а саме: класифікувати текстові масиви за результатами ЧЗЗ, що реалізовано в сучасних англомовних текстових редакторах типу Word for Windows.

1. Пиотровский Р.Г. Текст, машина, человек. Л., 1975. 2. Сеньківський В.М. Методологія проектування систем комп'ютерного підготування видань: Автореф. дис... на здобут. вчен. ступ. докт. техн. наук. Львів, 1966. 3. Шульмейстер М.В., Таль Г.А. Справочник технолога-полиграфиста. Ч. 1. Наборные процессы. М., 1981

Стаття надійшла до редколегії 27.01.97